Our Approach

We wanted to understand the data by segmenting the data into clusters

Factor Analysis Output

Exploratory Factor Analysis

**UPDATES ON THE PROGRESS MADE**

* From the KMO score we have come to a conclusion to include all 7 the variables in factor analysis

\* 0.00 to 0.49 unacceptable

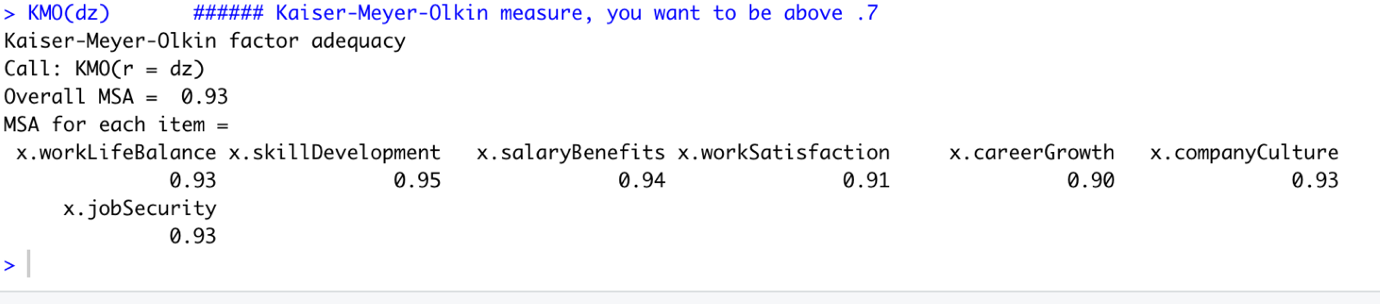
\* 0.50 to 0.59 miserable

\* 0.60 to 0.69 mediocre

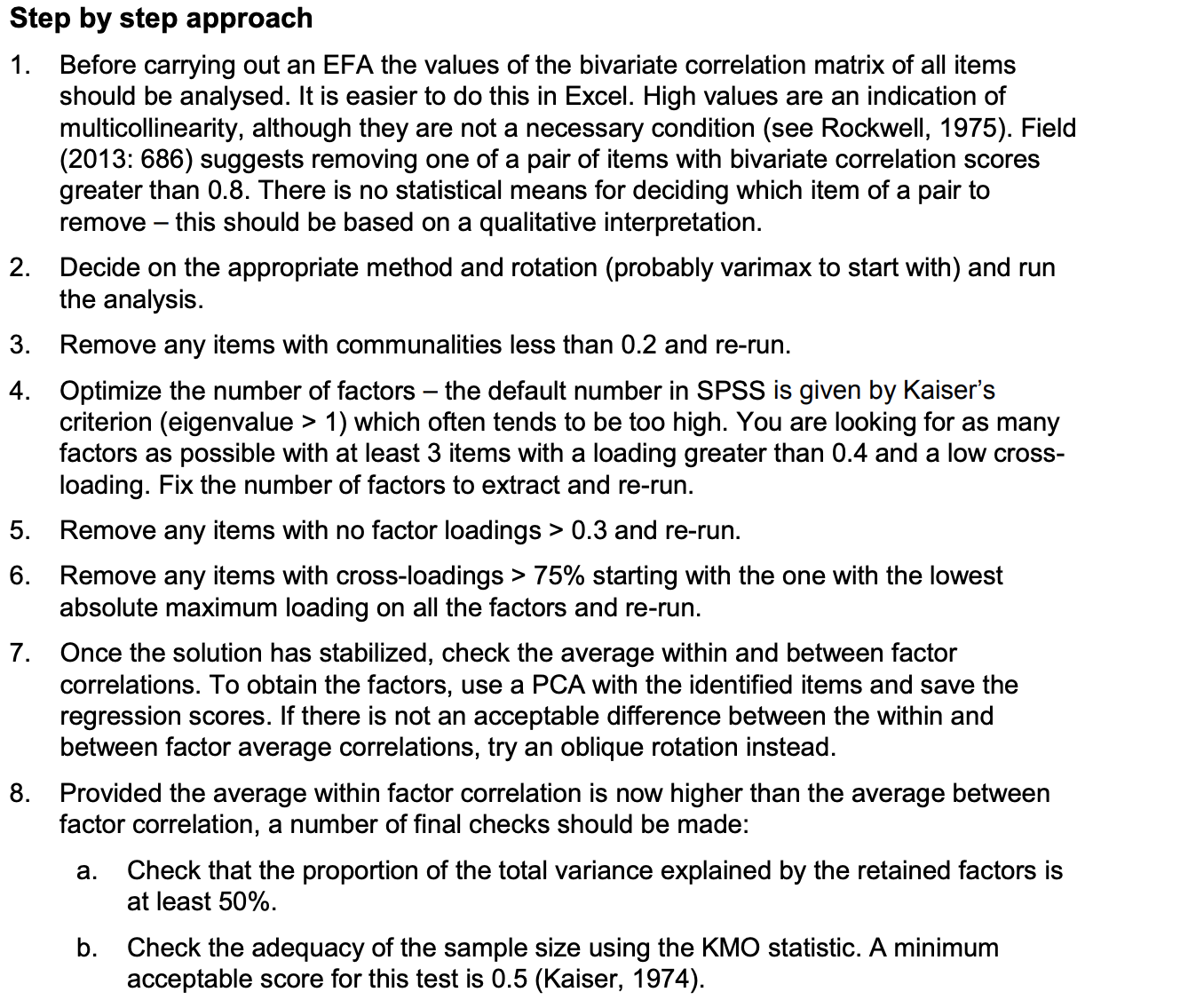
\* 0.70 to 0.79 middling

\* 0.80 to 0.89 meritorious

\* 0.90 to 1.00 marvelous



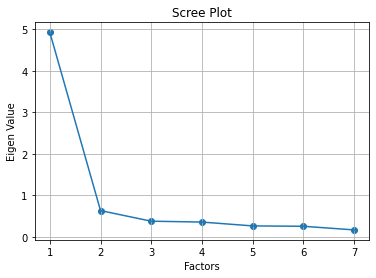
Using the below Research article, we could see that the number of **factors=2 or 1** are suitable for our data

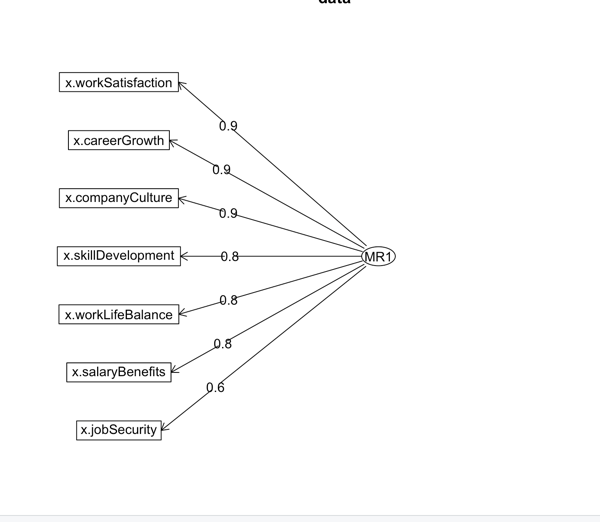
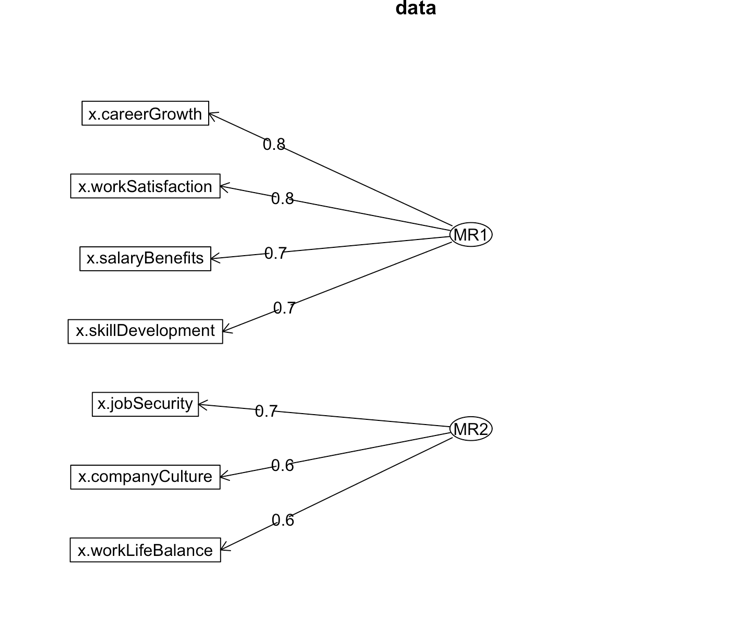


<http://www.open-access.bcu.ac.uk/6076/1/__staff_shares_storage%20500mb_Library_ID112668_Stats%20Advisory_New%20Statistics%20Workshops_18ExploratoryFactorAnalysis_ExploratoryFactorAnalysis4.pdf>

**Current Approach**

From the above Factor Analysis output we could see that one latent construct (factors) or two latent constructs from the 7 variables are good to have





We can interpret factor loadings just as we would interpret (Pearson) correlations.

0   
skillDevelopment -0.817145.    
salaryBenefits -0.765802.    
workLifeBalance -0.794560   
workSatisfaction -0.903469   
jobSecurity -0.636593   
careerGrowth -0.869812   
companyCulture -0.861450

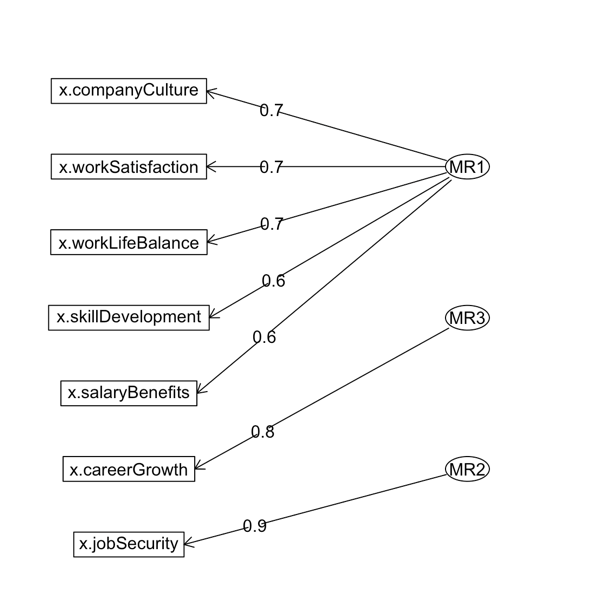
Communalities   
skillDevelopment 0.667726   
salaryBenefits 0.586452   
workLifeBalance 0.631325   
workSatisfaction 0.816256   
jobSecurity 0.405250   
careerGrowth 0.756573   
companyCulture 0.742095

Previous approach

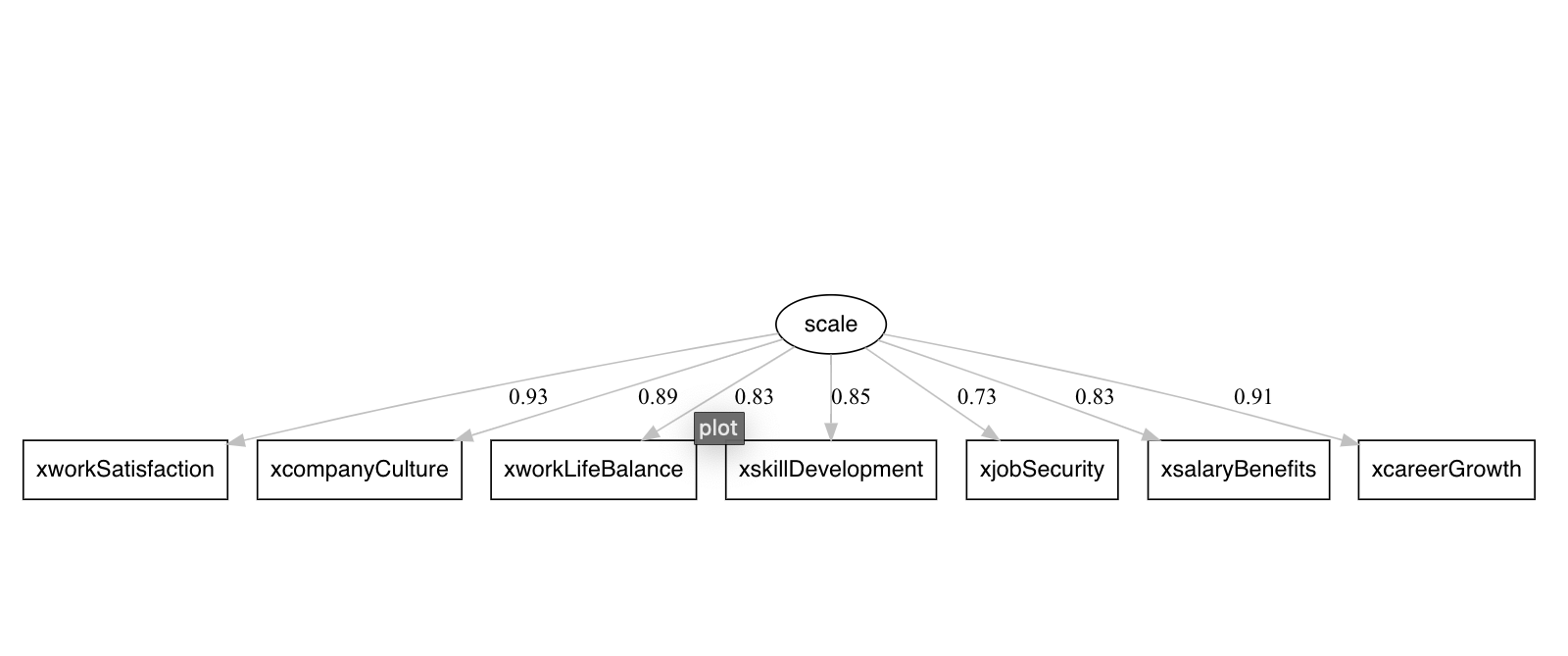
Theoretically, we could have as many factors as you do variables. But that, of course, defeats the purpose of running a factor analysis.

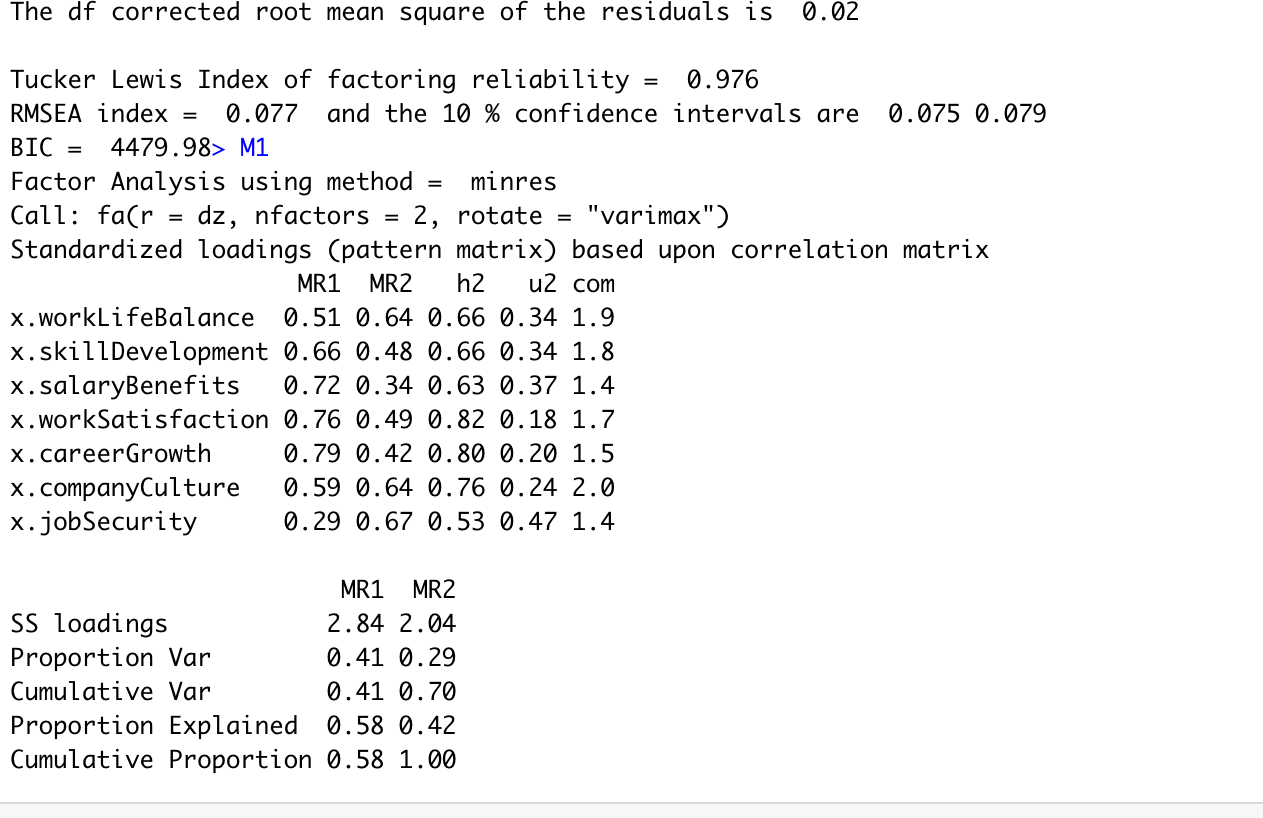
We were more inclined to get three factors to align our latent constructs with the defined burnout axis (exhaustion, depersonalization).

The feedback from the midreview echoes the same about  forcing the factors/clusters on data



Confirmatory Factor Analysis also validating taking two or one factor sufficient

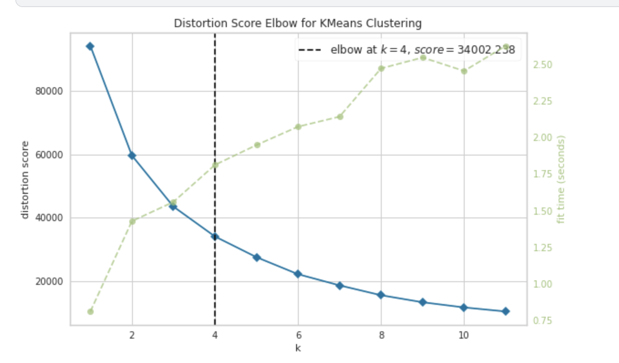


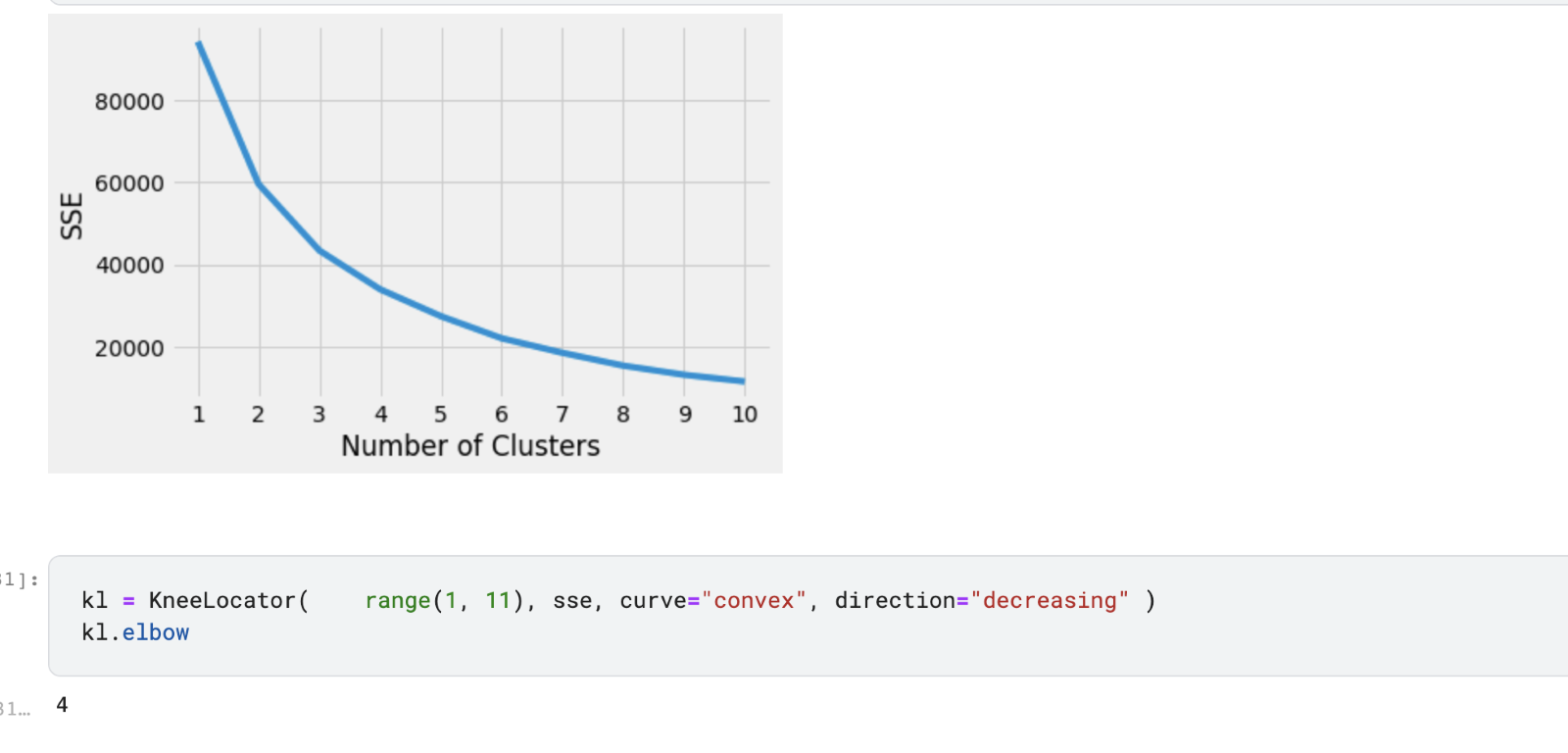


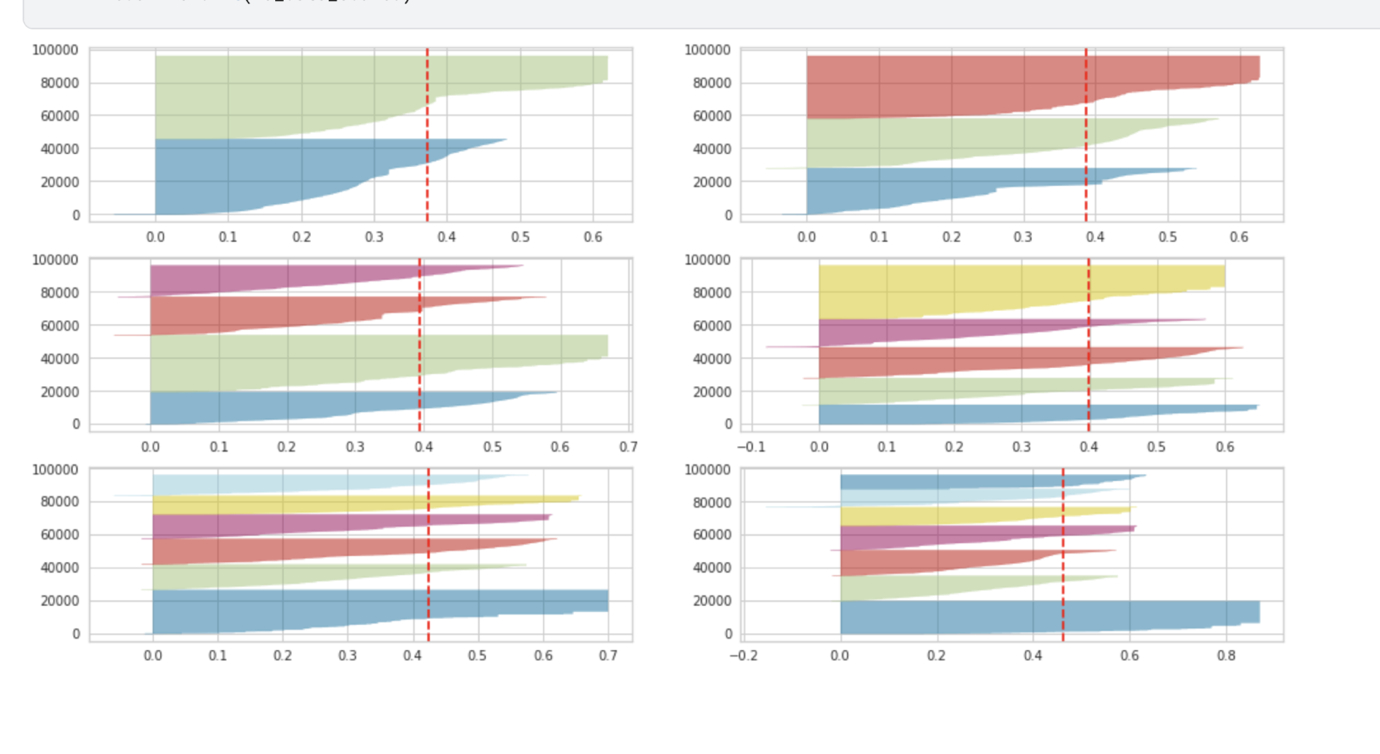
**Cluster analysis post the factor Analysis**

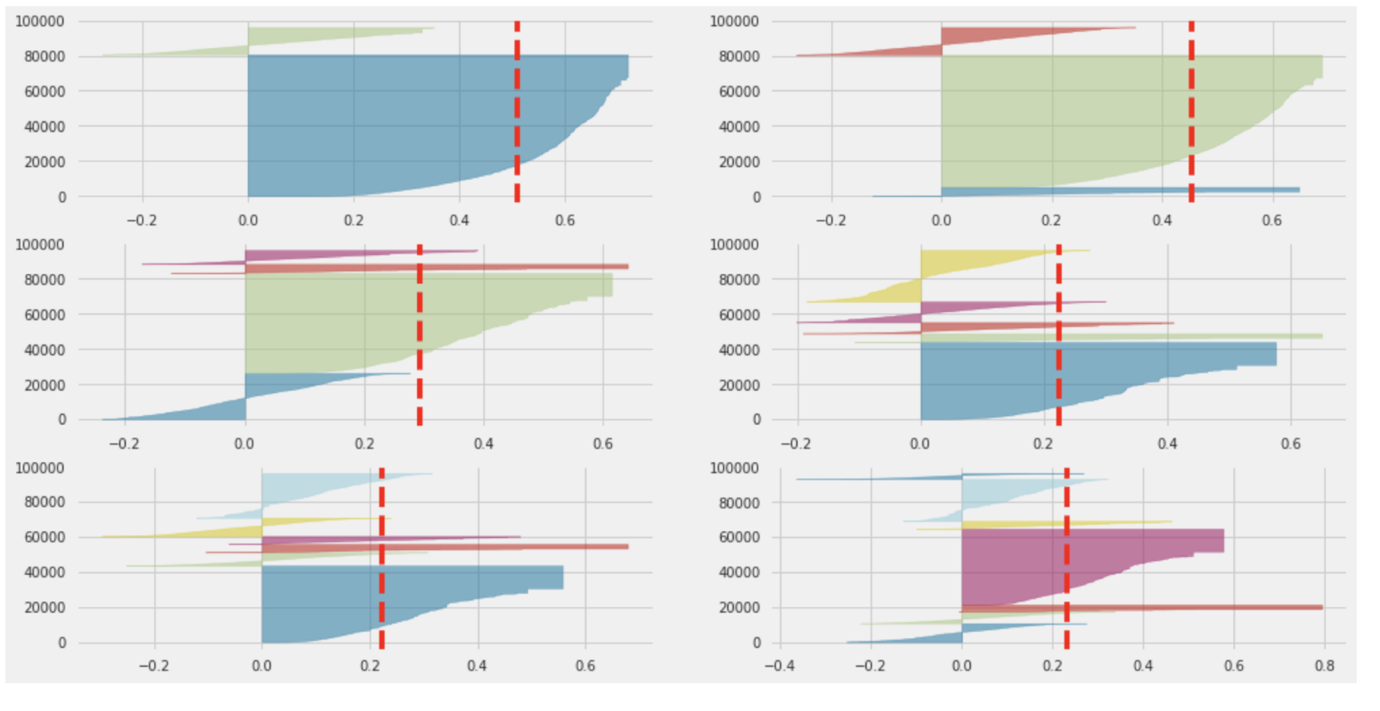
As per our discussions in the previous call we used clustering on the factors of the data

Finding out the best number of k for clustering from elbow plot and silhouette plot



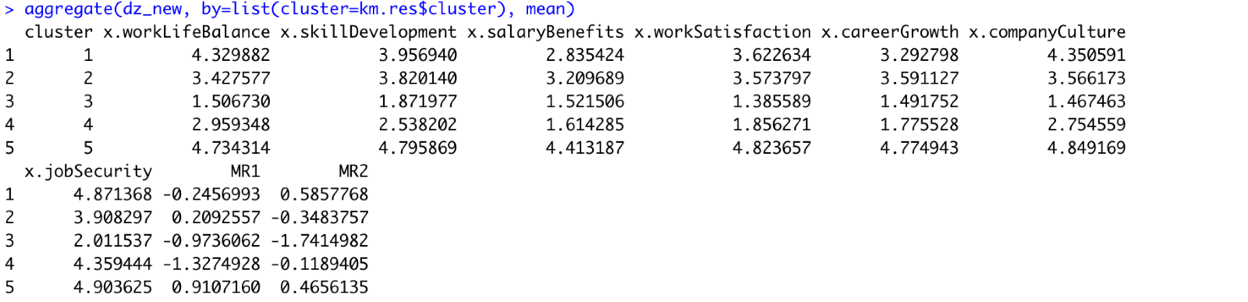


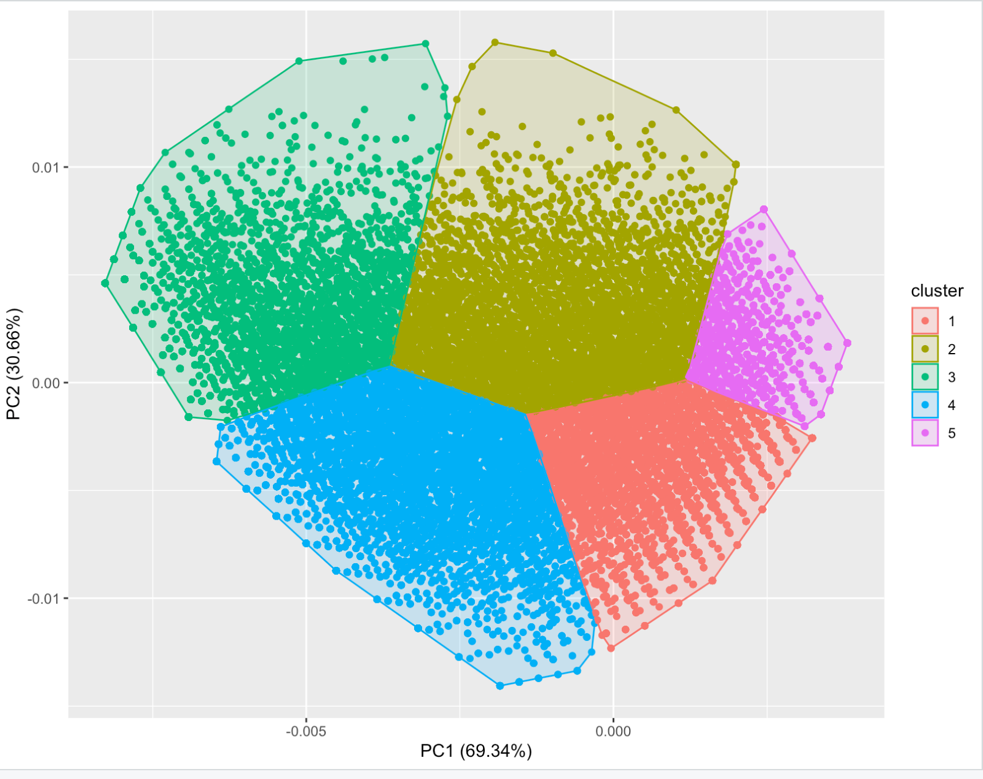




**2 factors and 5 clusters** /**3 clusters**

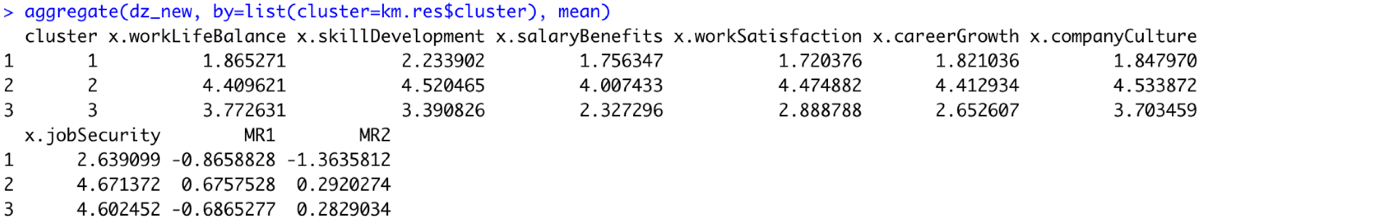
**(The better clustering out of all the experiments done)**





2 factors and 3 clusters





Having factors more than 2 resulted in poor clustering of the data

Table

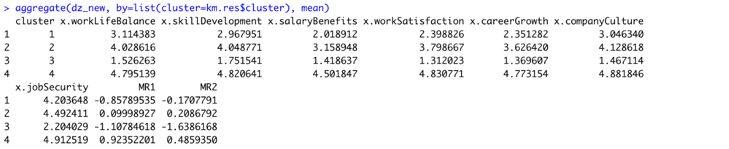
Description automatically generated with medium confidence

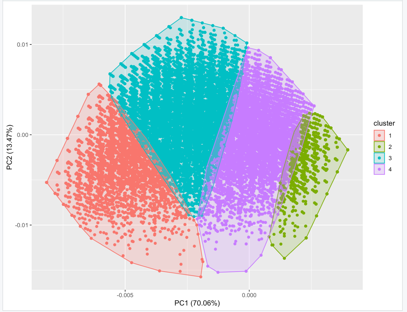
Chart, surface chart

Description automatically generated

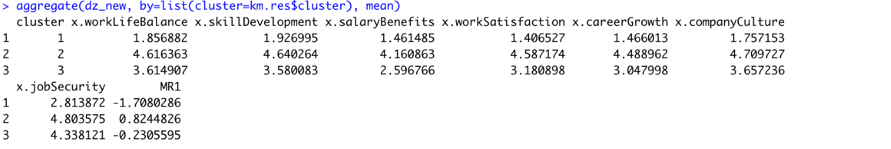
Screenshots of the experiments we rejected out due to overlapping of the segments

2 factors and 4 clusters (overlapping is present)



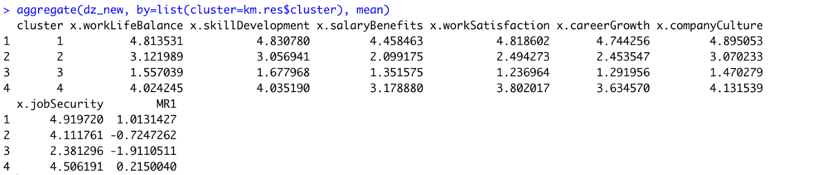


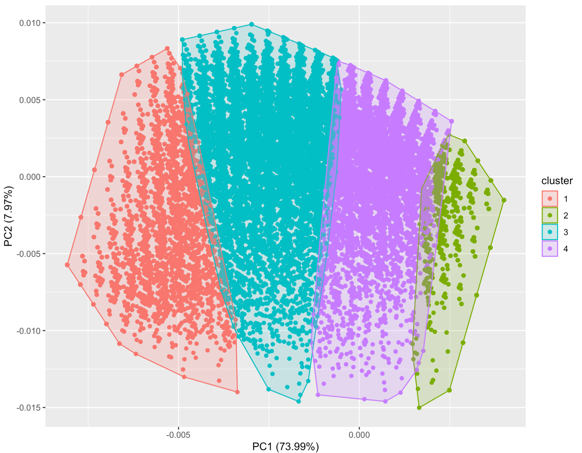
One factor and 3 clusters )overlapping is present



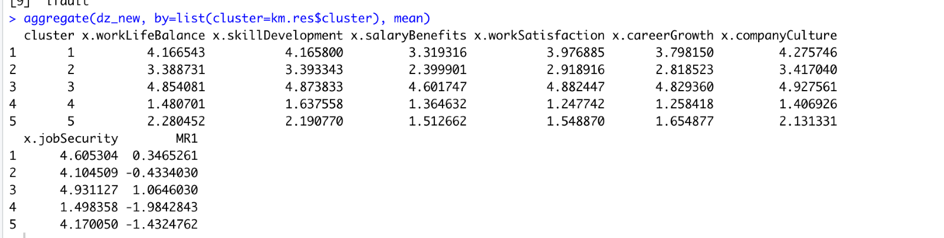


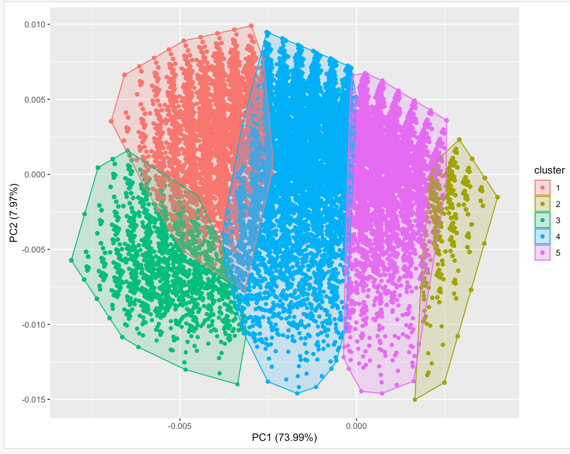
One factor 4 clusters (overlapping is present)





One factor 5 clusters (overlapping is present)





FUTURE STEPS

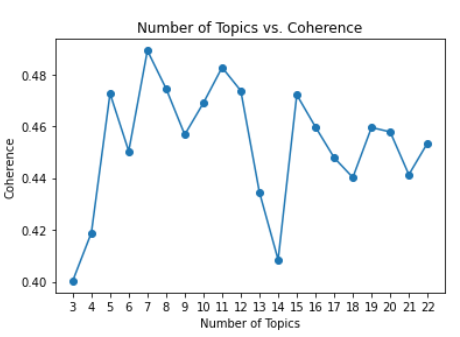
* Given our approach of clustering into 3 segments and the factor analysis .we would like to proceed using the clustered category text for further analysis.

(Ex.Analysing cluster 1 text and topic modelling of cluster one helps us in understanding the sentiment of the people who have given very less ratings to the company )

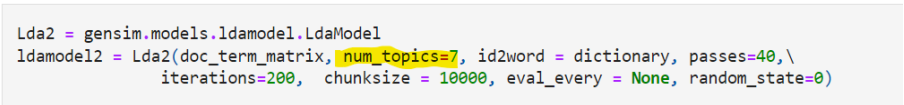
* We would like to translate this to a recommendation system or a Multiclass classification approach (As we have data that is labelled into clusters now)
* Converting the text to a sentiment score and validating our approach with text and without text scores

Fine Tuning LDA:

LDA requires that we specify the number of topics that exists in a corpus of text. There are several common measures that can be optimized, such as predictive likelihood, perplexity, and coherence. Much literature has indicated that maximizing coherence, particularly a measure named Coherence Value  (https://svn.aksw.org/papers/2015/WSDM\_Topic\_Evaluation/public.pdf), leads to better human interpretability. This measure assesses the interpretability of topics given the set of words in generated topics. Therefore, we will optimize this measure. Result of Coherence analysis on text corpus



LDA Model Selected basis the ‘CONS’ column:



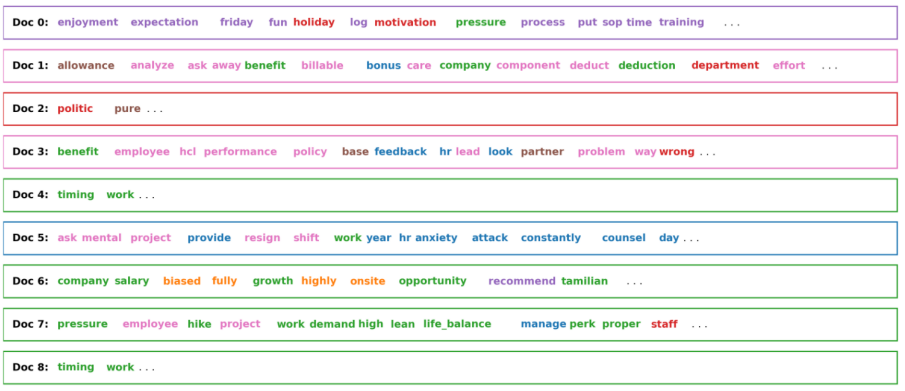
LDA Output basis the ‘CONS’ column:



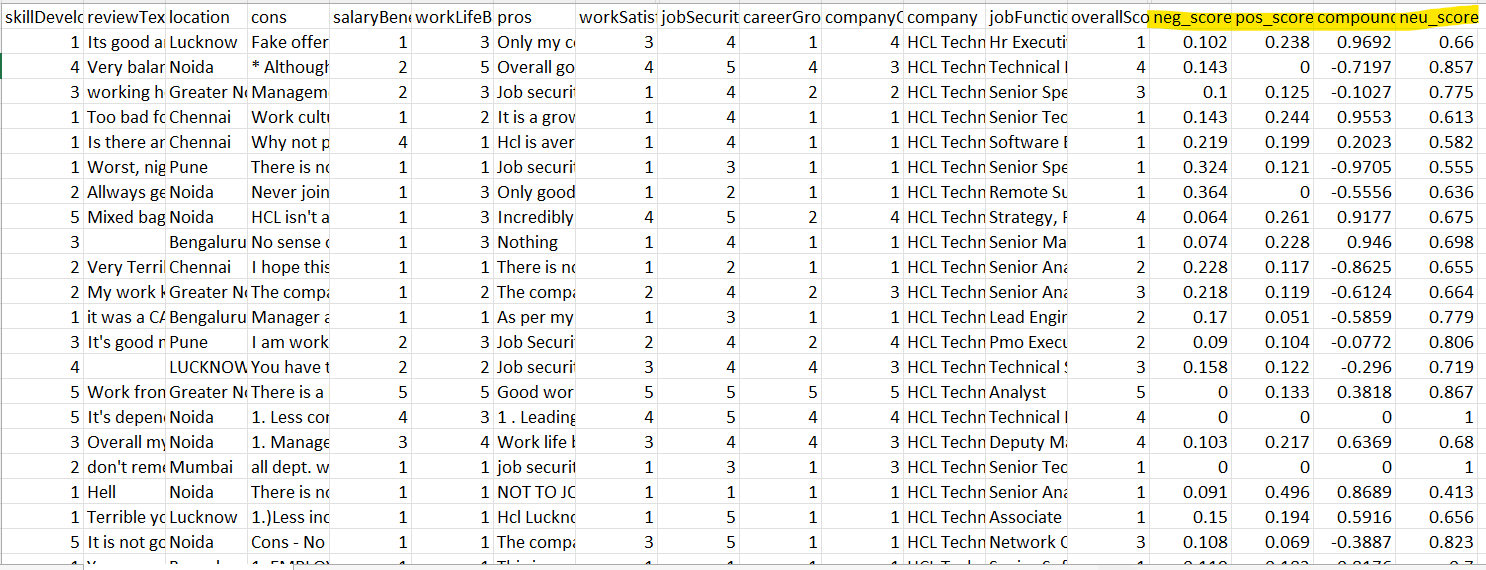
* Need to name these topics on basis of subjective and domain expertise

LDA Current State and Next Steps

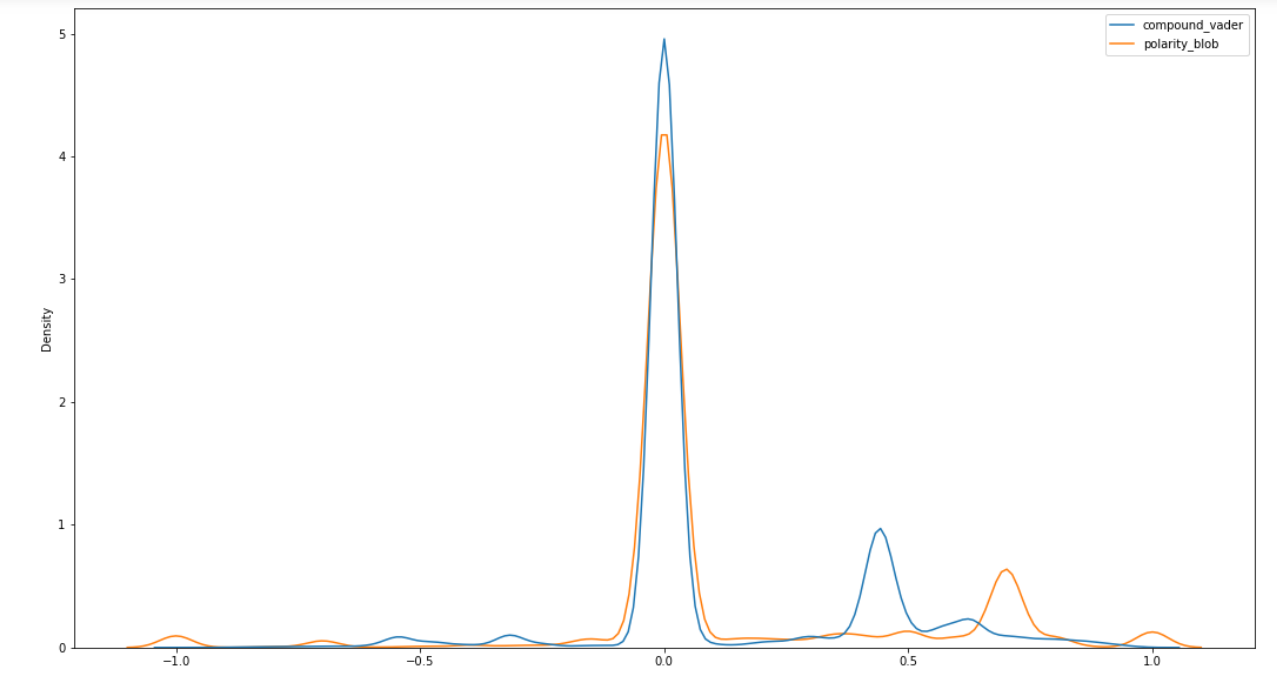
• We were able to split each review by topic



* Extract the most dominant topic in each review text and perform sentiment analysis to get a final sentiment score



* Fine-tuning the sentiment analysis results
* Tried TextBlob and VADER to arrive at the sentiment score on ‘REVIEW TEXT’



* Sentiment score can be used as additional feature. It will also lay foundation to Suggestion Box, Water Cooler etc. – Next Steps

Articles referred to estimate the factor scores

<https://www.researchgate.net/publication/255643537_Understanding_and_Using_Factor_Scores_Considerations_for_the_Applied_Researcher>

<https://medium.com/@Gabrielefrattini/clustering-market-segments-with-k-means-and-factor-analysis-9b3876581b16>

<https://rpubs.com/pjmurphy/758265>

<https://github.com/Dhamodaran-Babu/Machine-Learning-Exercises/tree/master/12.Factor%20Analysis>

<https://www.tqmp.org/RegularArticles/vol09-2/p079/p079.pdf>